Data Report:
Predicting Baseball Players' Salaries

Overview

Baseball_Data_Exploration_and_Preprocessing notebook is the first step in data exploration by importing appropriate files and dataframes. These dataframes were analyzed by assessing linear regressions and visualizations. The notebook focuses on the objective question: How can salaries of baseball players be predicted based on their game performance and popularity?

1. One potential actionable is to help youth evaluate themselves based on categories that will estimate their salary.
2. The second actionable is to help team managers better understand their players' salaries to prevent overpaying or underpaying.
3. The last actionable is to help baseball players have better understanding of their individual records to decide if they would like to move or stay in their team.

The Initial_Visualizations notebook visualizes potential relationships between columns in baseball data. For example, we explored the relationship between Batting, Pitching, HallOfFame, AwardsSharePlayers, and salary. The relationship between fielding and salary proved to be the strongest.

OLS_regression_results notebook performs linear regression analysis and fits a best-fit line on a scatter plot to predict the dependent variable based on one or multiple independent variables. The best-fit line represents the best model that minimizes absolute error from the dispersed scatter point.

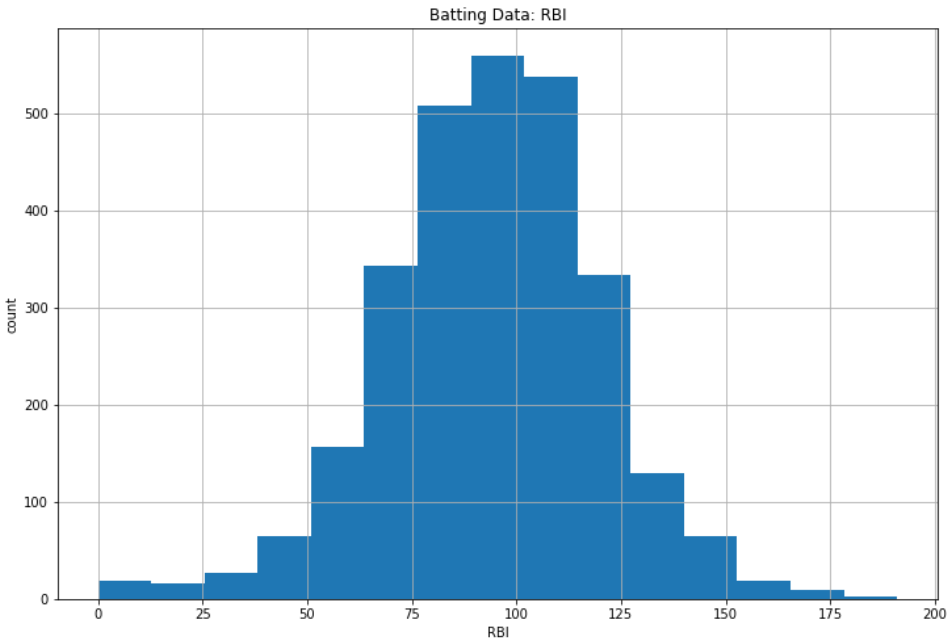
The report will discuss the focus problem and introduce background information on baseball data for analysis interpretation.We will explain potential features with visualizations that show relationships between baseball data and salaries. Linear regressions analysis will analyze the best-fit line that minimizes absolute error based on the context of baseball. Then, insights will be drawn to provide a reasonable conclusion to the final data report.

Background

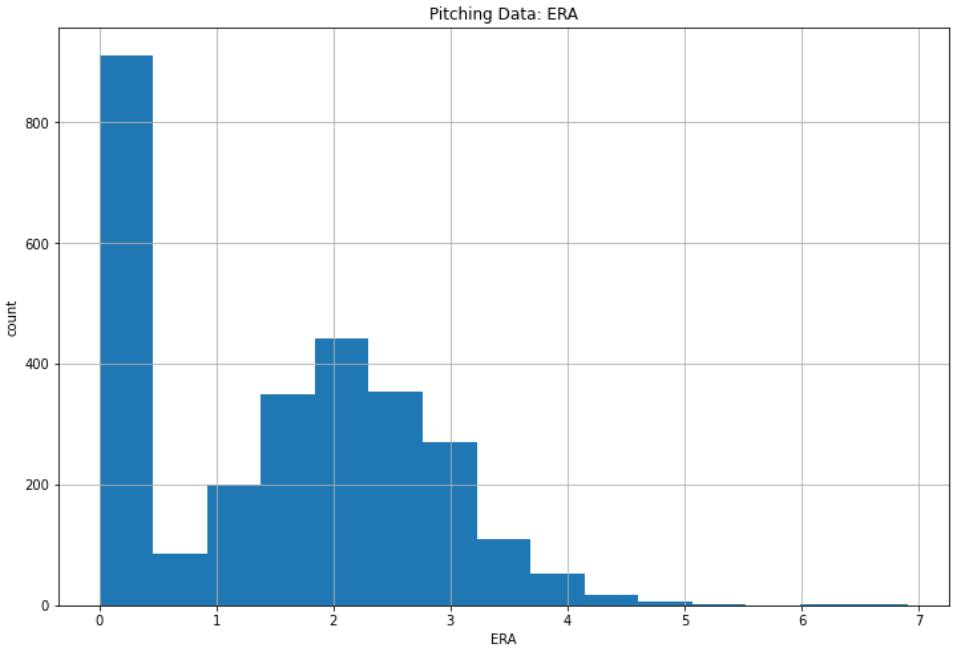
The focus question of the data report is: How can we predict salaries of baseball players based on their game performance and popularity? Appropriate files were chosen from the Baseball Databank found in https://www.kaggle.com/datasets/open-source-sports/baseball-databank. AwardsSharePlayers.csv, Batting.csv, Fielding.csv, Pitching.csv, HallOfFame.csv, Teams.csv, and TeamsHalf.csv were uploaded to data exploration report to merge with Salaries.csv and

pointsWon column was located from AwardsSharePlayers.csv, which is the number of votes to nominate the award winner. RBI (Runs Batted In) records are quantitative data that add up every time a hitter scores a point in the game. The official definition is provided by MLB.com, "A batter is credited with an RBI in most cases where the result of his plate appearance is a run being scored." RBI was collected from Batting.csv. Error records located from Fielding.csv add up every time a defender misses to field a play. ERA (Earned Run Average) is a pitching data from pitching.csv that increases whenever pitchers show a bad performance in the game. The official definition is provided by MLB.com, "Earned run average represents the number of earned runs a pitcher allows per nine innings—with earned runs being any runs that scored without the aid of an error or a passed ball." Ballots from HallOfFame.csv follow the similar concept with pointsWon from AwardsSharePlayers.csv to be nominated into the Hall of Fame. The data on team wins in Teams.csv and TeamsHalf.csv is the number of wins in a single season (total 162 games). Histograms and scatter plots visually represented the relationships between RBI, ERA, Wins, Ballots, Errors, and Salaries to answer the focus question. With reasonable data and initial analysis, we moved onto the second part of the project, which was to predict salaries of baseball players using a best-fit line.

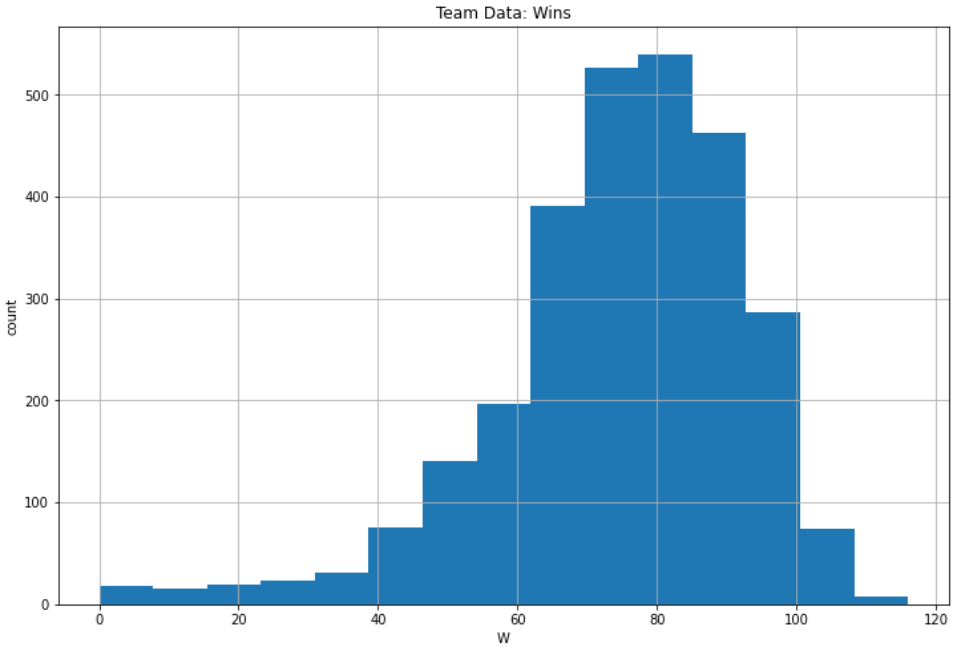
Potential Features



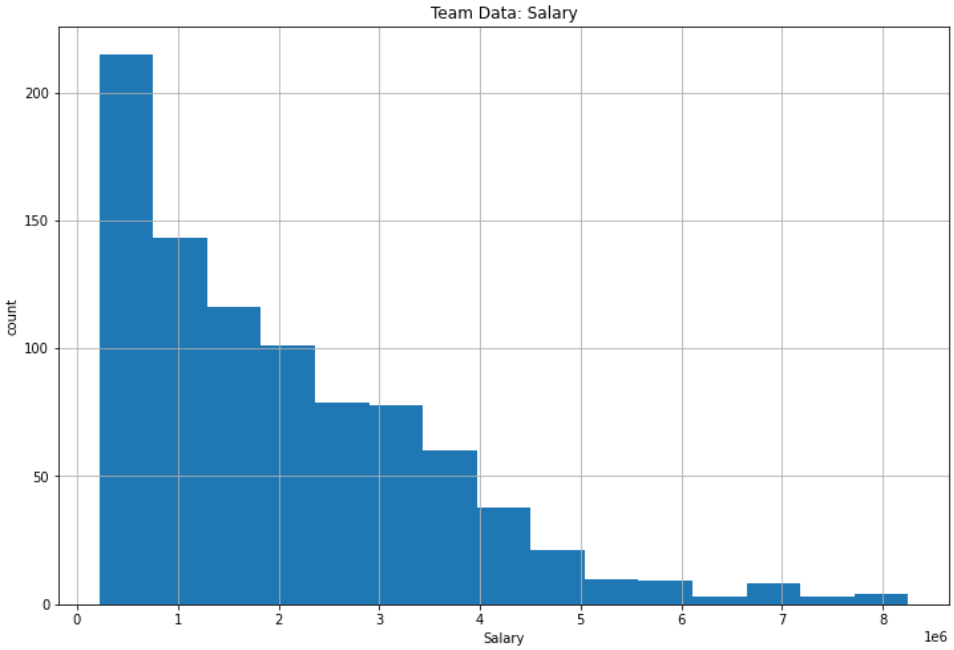
This graph shows the hitters' RBI records. RBI is a potential feature because the data represents how many times a hitter can make a runner score, which is an accurate reflection of their hitting performance in games. The maximum RBI is close to 190 and the minimum RBI is 0 in a single season based on the histogram. The average RBI for players seems to be about 80 to 90, which is a reasonable amount.



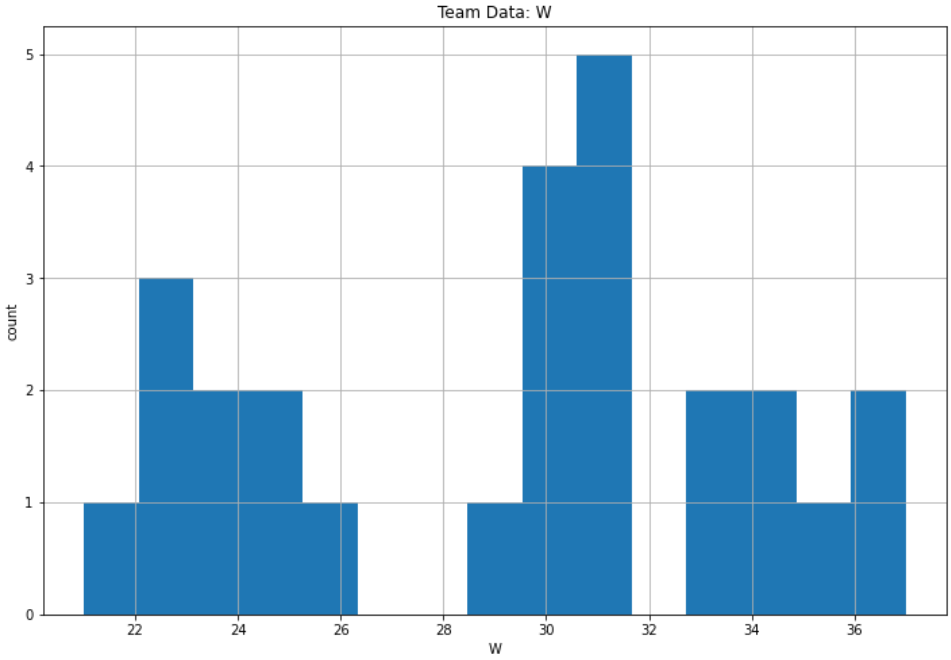
This graph shows the pitchers' ERA. ERA is a potential data because it represents average numbers of hits that pitchers allow when they enter the game. If a pitcher has high ERA, this means that the player is not capable of disallowing hits. Since the x-axis of the histogram is ranging from 0 to 7, the dataset is verified considering that there aren't a lot of ERA records over 7. The minimum value of ERA in a single season was 0.00, while the maximum value was around 7.00.



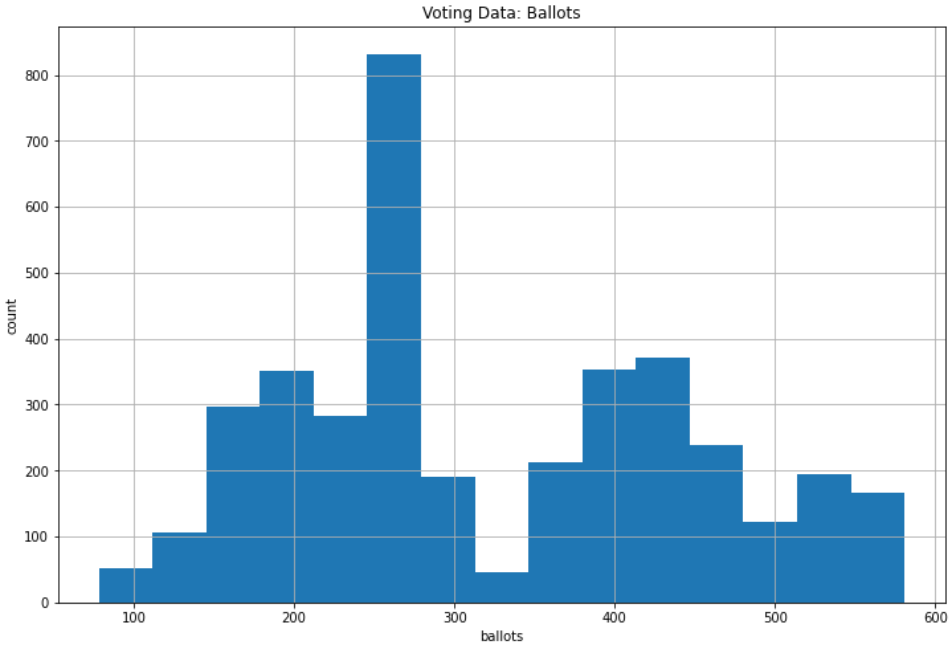
The graph shows the number of wins of all baseball teams in a single season with a maximum of around 118 wins and at least 0 wins. Since there are about 160 games in a single baseball season, the maximum number of wins seems to be fair considering that the winning percentage was about 72.8%. The average number of wins sticks around 70 to 80 wins, which is again reasonable because the mean winning percentage should be around 50%.



The graph shows the average salary of baseball teams' they invest on their players. Less concentration in this graph because starplayers are spread out in different teams to represent that team as a franchise. Teams in 7~8 range have high average salary, but this does not necessarily mean that the total salary is high because there is a possibility that the number of players is different for every team.

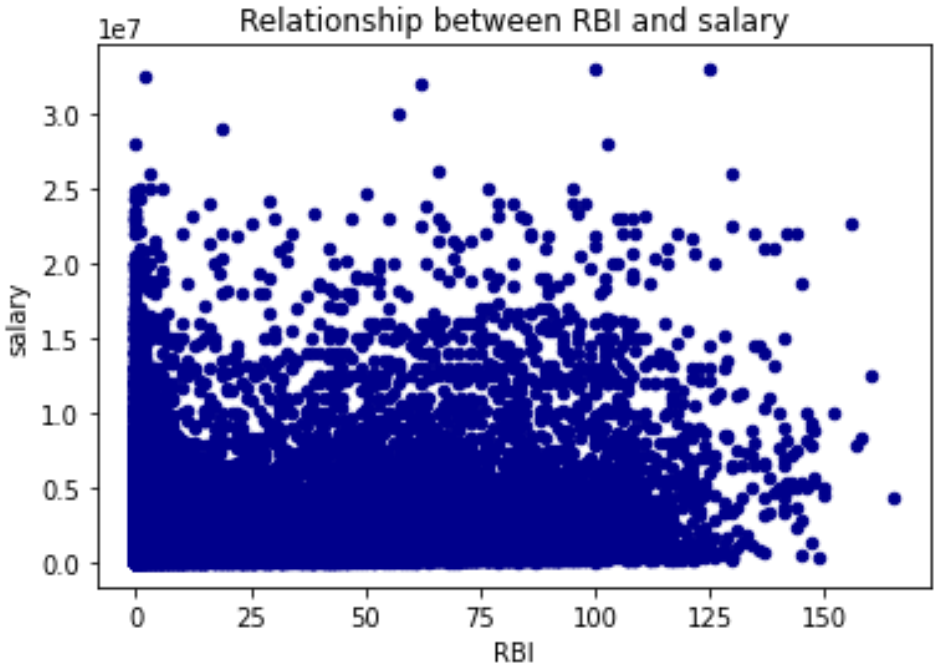


The graph shows the number of half wins for different baseball seasons. The graph is fairly distributed because half of a season is less games for teams to increase or decrease the gap between teams.

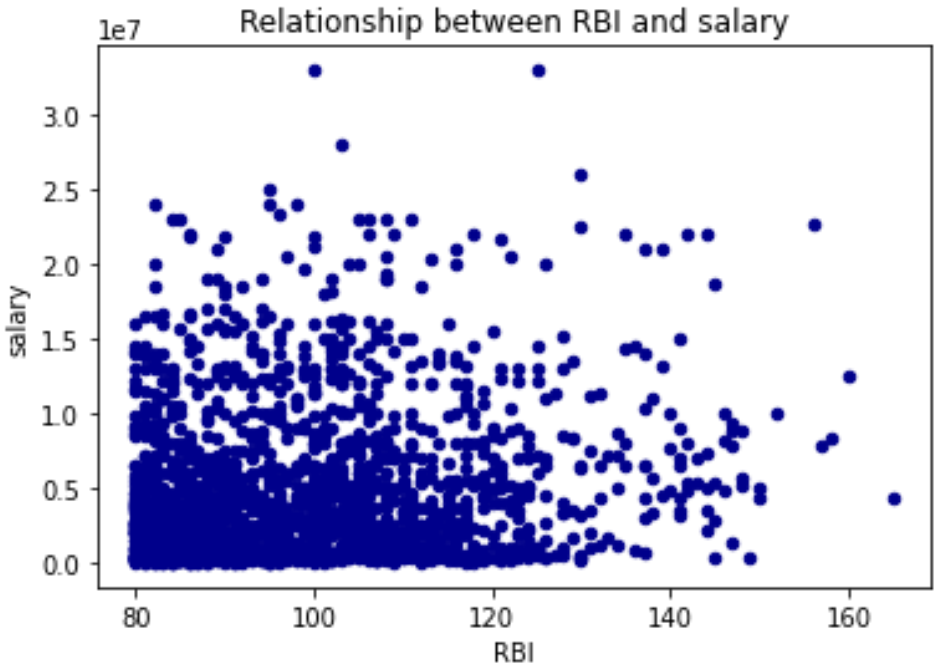


The graph shows the number of ballots a player received to be designated a place in the Hall of Fame. The average number of votes rested around 300 to 400, while the maximum number of ballots ever received was over almost 600. This histogram is more evenly distributed than other histograms because the players listed in the dataframe are possible candidates for an honored

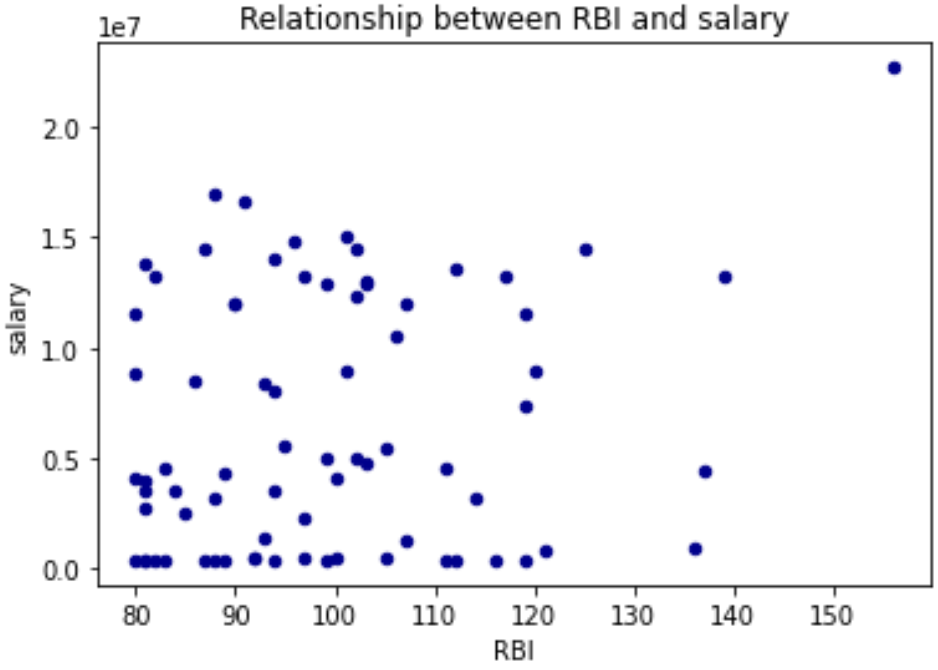
recognition. This means that the players listed in the dataset are already popular players, so players with less popularity were not included in the candidate selection.



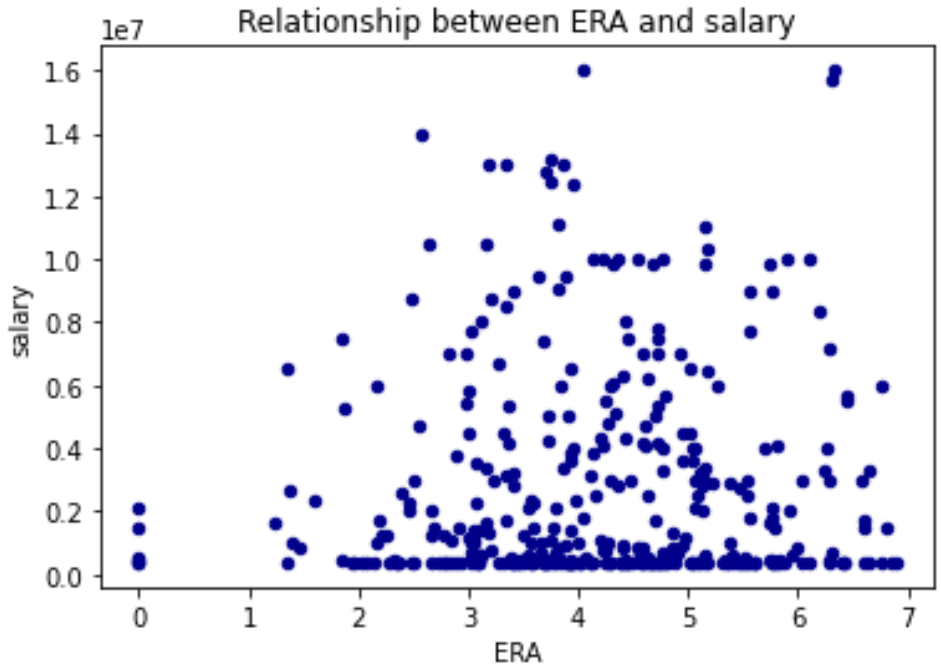
Players may be more famous for fielding than hitting, players may be more famous for pitching than hitting. Players' salaries are based on players' performance in the last season (also why players with high RBI have low salary). Inflation in the years past may be another factor that can make this data inaccurate in historical context. The scatterplot was far from the expected result because the plot shows low RBI players with high salary. The reason for the trend of data points in low RBI records but high salary range is because there are players who qualify for other baseball skills than batting.



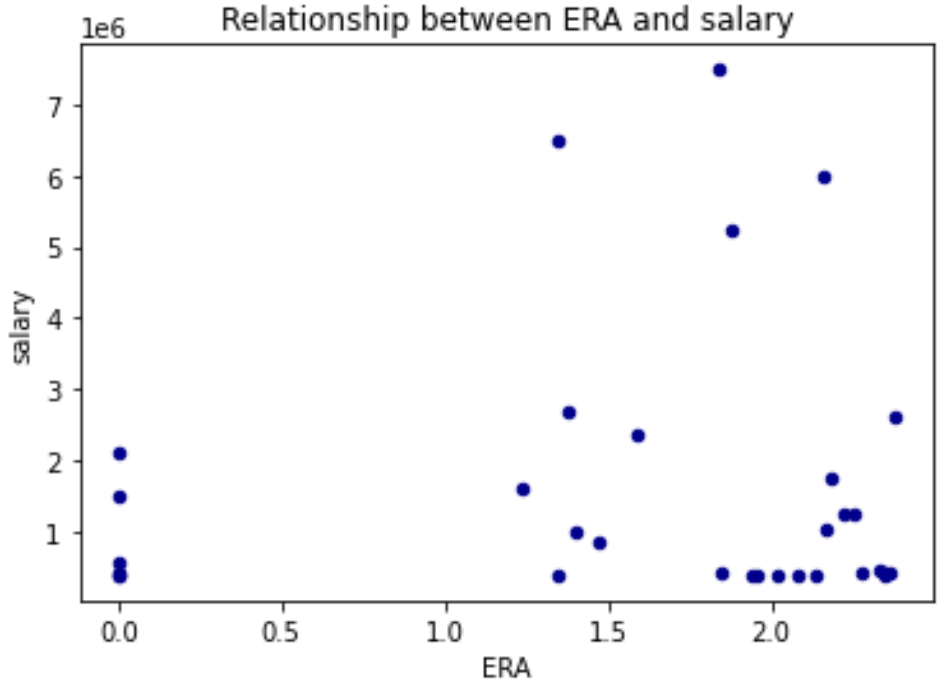
Approach: cut/filter players with almost 0 RBI because they should be good at something else use previous RBI table (25%) to get top 75% players.



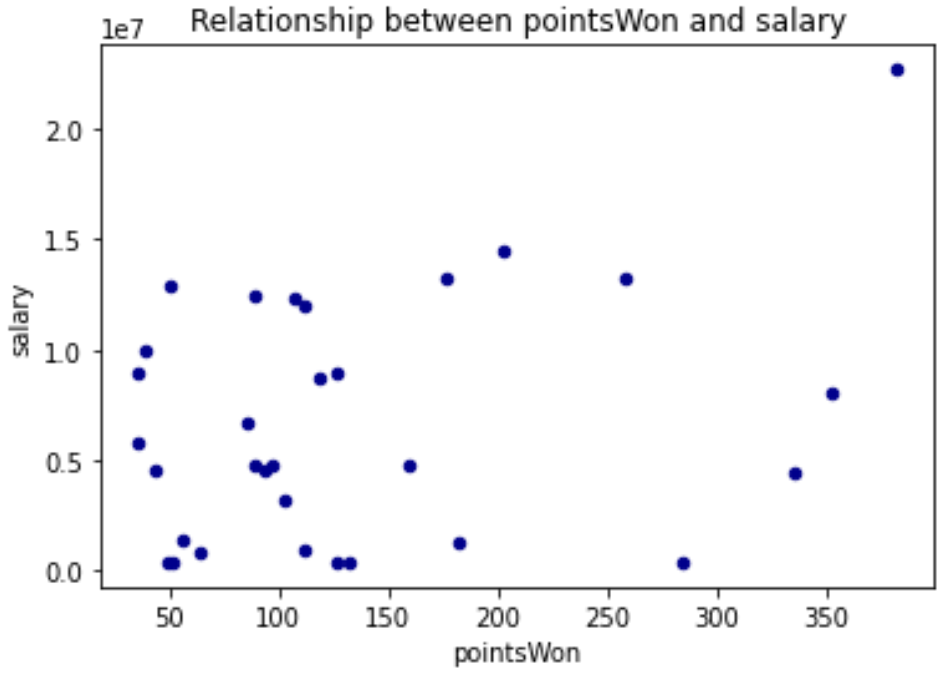
This scatterplot is even closer to the expected trend than the previous scatter plot because there seems to be an exponential growth in the relationship between RBI and salary. The relationship between the two datasets exists, but is subtle. There may be other factors that more significantly influence a player's salary.



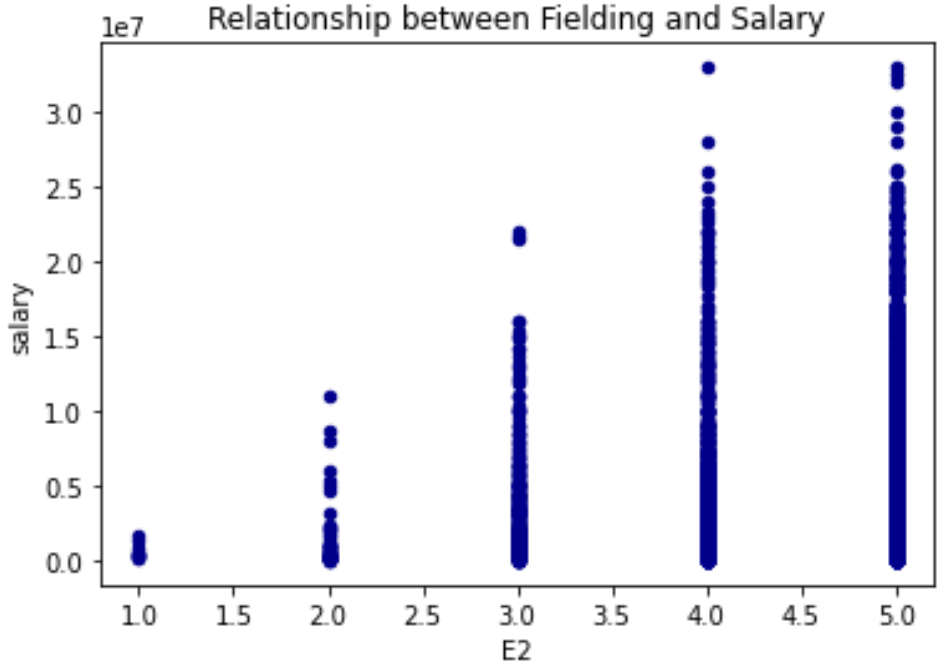
The scatterplot shows the relationship between ERA and salary to be a slightly positive correlation. The result is far from the expected result because a low ERA is good pitching data, but it seems that the salary is lower for good ERA pitchers. This correlation is likely due to ERA fluctuating a lot by minor mistakes in the game, which is the reason why ERA below 2 or 3 are considered generally good.



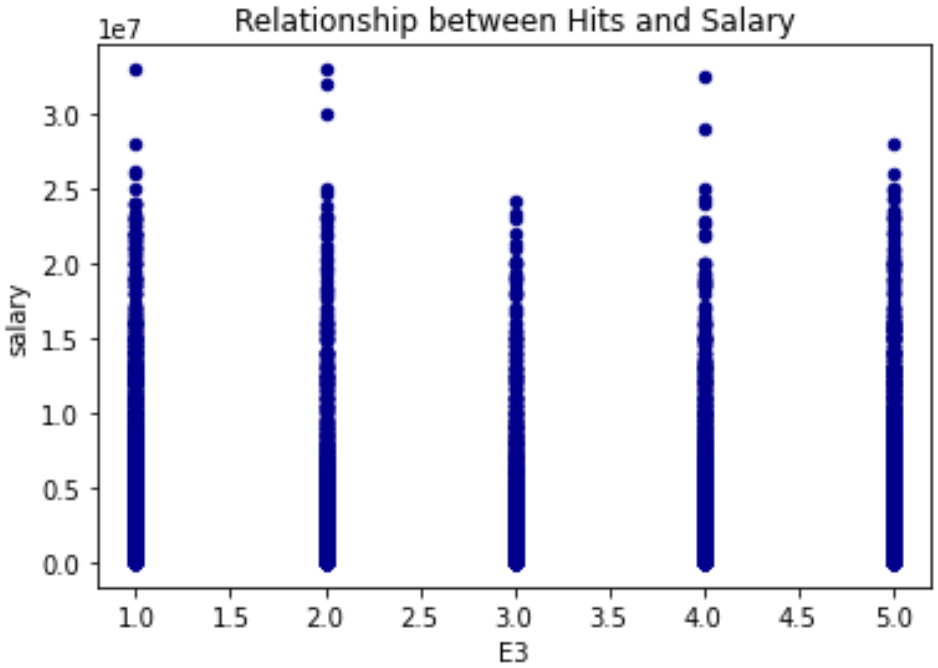
This merged graph shows the relationship between pitchers' ERA and salary. The graph is not really close to the expected correlation because ERA fluctuates a lot every year for players, so the data that year may not properly reflect the player's normal performance.



This scatterplot is closer to the prior expected result because the pointsWon and salary seems to have somewhat of a positive correlation. This poses a possible relationship between popularity and salary in baseball.



This scatterplot shows that good fielding isn't enough to guarantee a high salary. However, a bad rating at fielding, which would be around the range of 1 to 3, will bring players less salary compared to players with high fielding rates.



This plot does not suggest any meaningful trends.

Linear Regressions Analysis

Linear regression analysis draws a best-fit line on a scatter plot to predict the dependent variable based on one or multiple independent variables. The best-fit line represents the best model that minimizes absolute error from the dispersed scatter point. These are the linear regression results and interpretations from my OLS_regression_results notebook:

OLS Regression Results

Dep. Variable:	salary	**R-squared:**	0.041
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	5.694
Date:	Thu, 24 Nov 2022	**Prob (F-statistic):**	0.00379
Time:	21:20:30	**Log-Likelihood:**	-4630.7
No. Observations:	270	**AIC:**	9267.
Df Residuals:	267	**BIC:**	9278.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-1.583e+06	3.07e+06	-0.516	0.606	-7.62e+06	4.45e+06
pointsWon	-7329.4212	4059.274	-1.806	0.072	-1.53e+04	662.838
RBI	1.036e+05	3.11e+04	3.329	0.001	4.23e+04	1.65e+05

Omnibus:	21.081	**Durbin-Watson:**	1.539
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24.221
Skew:	0.732	**Prob(JB):**	5.50e-06
Kurtosis:	3.087	**Cond. No.**	1.24e+03

The R-squared value is 0.041, which is not a strong correlation because it is not close to 1. The number of observations is 270, which was improved by applying a wider range of yearID to show up in the dataset. The p-values for pointsWon and RBI are below 0.05, which add direction of correlation.

OLS Regression Results

Dep. Variable:	salary	**R-squared:**	0.023
Model:	OLS	**Adj. R-squared:**	0.014
Method:	Least Squares	**F-statistic:**	2.620
Date:	Thu, 24 Nov 2022	**Prob (F-statistic):**	0.0750
Time:	17:04:58	**Log-Likelihood:**	-3915.6
No. Observations:	227	**AIC:**	7837.
Df Residuals:	224	**BIC:**	7847.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.006e+07	9.33e+05	10.779	0.000	8.22e+06	1.19e+07
pointsWon	-1.23e+04	5415.219	-2.272	0.024	-2.3e+04	-1631.497
E	6.76e+04	1.13e+05	0.599	0.550	-1.55e+05	2.9e+05

Omnibus:	15.222	**Durbin-Watson:**	1.109
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16.064
Skew:	0.618	**Prob(JB):**	0.000325
Kurtosis:	2.590	**Cond. No.**	301.

The R-squared value is 0.023, which is not a strong correlation because it is not close to 1. The number of observations was initially 36, but was improved to become 227 observations after a great range of yearID was applied. Coefficient of variables shows that the correlation between two independent variables and dependent variable is slightly positive. Regression results show that the linear regression is not significant because we have a variable E with a high p-value.

OLS Regression Results

Dep. Variable:	salary	**R-squared:**	0.242
Model:	OLS	**Adj. R-squared:**	0.242
Method:	Least Squares	**F-statistic:**	385.8
Date:	Thu, 24 Nov 2022	**Prob (F-statistic):**	4.08e-146
Time:	19:36:38	**Log-Likelihood:**	-40514.
No. Observations:	2417	**AIC:**	8.103e+04
Df Residuals:	2414	**BIC:**	8.105e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-5.546e+08	2.1e+07	-26.434	0.000	-5.96e+08	-5.13e+08
pointsWon	8121.3685	1146.295	7.085	0.000	5873.544	1.04e+04
yearID	2.792e+05	1.05e+04	26.620	0.000	2.59e+05	3e+05

Omnibus:	529.555	**Durbin-Watson:**	1.547
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1360.544
Skew:	1.173	**Prob(JB):**	3.64e-296
Kurtosis:	5.830	**Cond. No.**	4.48e+05

The p-values are strong because they are less than 0.01 (99% confident). We are confident that the variables are correlated with salary; the R-squared value suggests that this linear regression explains the most about variance in salary. The R-squared value is far from 1 because there are lots of factors that explain the variance in salaries, but not captured in our data. Other factors include cost of living, leadership, experience, past teams.

OLS Regression Results

Dep. Variable:	salary	**R-squared:**	0.031
Model:	OLS	**Adj. R-squared:**	0.030
Method:	Least Squares	**F-statistic:**	30.47
Date:	Thu, 24 Nov 2022	**Prob (F-statistic):**	9.44e-14
Time:	19:35:16	**Log-Likelihood:**	-31854.
No. Observations:	1892	**AIC:**	6.371e+04
Df Residuals:	1889	**BIC:**	6.373e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.327e+06	3.47e+05	6.712	0.000	1.65e+06	3.01e+06
pointsWon	6983.2764	1347.861	5.181	0.000	4339.823	9626.730
G	1.204e+04	2755.891	4.368	0.000	6632.258	1.74e+04

Omnibus:	606.617	**Durbin-Watson:**	1.296
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1677.682
Skew:	1.688	**Prob(JB):**	0.00
Kurtosis:	6.145	**Cond. No.**	461.

The R-squared value is 0.031, which is not a strong correlation because it is not close to 1. But this is better than the previous regression results because R-squared has slightly increased. All P values are below 0.01, which make the correlation 99% confident. This was different from the other regression results because this regression only has variables that are statistically significant.

Data Insights

Actionables were identified prior to data research, and these are answered based on the data information collected. The findings conclude that a players' popularity and votes are more significant factors than batting, fielding, or pitching records to their salary.

1. Youth baseball players can be evaluated on categories that matter in a form of potential salary estimator.
   The focus of this research was to figure out significant categories that affect a player's salary and apply this knowledge to youth players. This kind of salary estimator would hopefully help youth players to find out their weaknesses and know what they have to work on. Unfortunately, this action cannot be accomplished from the data results because the relationships between salaries and particular baseball skills aren't strongly correlated to yield a reasonable prediction. Considering that there were negative salaries when predicted, the data results prove not to be applicable for usage as a salary estimator.

2. Team managers can have a better idea of their players' salaries whether they are being overpaid or underpaid.

Through the prediction of salaries, we hoped to develop a system that would help team managers to have a better understanding on which players are getting overpaid and underpaid. But due to the similar reason as the first actionable, prediction isn't applicable to the final data sets. Though a hard judgment is hard, team managers can view the data provided and refer to some relationships to have a better idea of their players' salaries. We can inform the team managers that players are paid more by popularity than skills. This will incentivize them to find more undervalued players who might be good at baseball but unpopular so far.

3. Baseball players can have a better idea of their salaries to choose whether they should move or stay in the team.
   This is similar to the second actionable because only the audience has changed. The dataset can also help baseball players to choose whether they should stay or move their team based on their salary and expected salary. Though a hard judgment is hard, baseball players can view the data provided and refer to some relationships to have a better idea of their salaries. We can suggest players who want more money to not only practice their baseball skills, but also to focus on their popularity.

Conclusion

The steps to data research on Baseball Dataset included data exploration, preprocessing, linear regression modeling, and visualization modeling. I was able to visually represent the relationships between several baseball data (RBI, ballots, pointsWon, G, ERA, E …) and salary to answer my focus question. Identification of statistically significant features was another key step in predicting players' salary in a reasonable and statistically proven manner.

There were also limitations to this data research project. I lost a lot of data points because datasets had different players and years, which made it difficult to find several intersection points. This made the number of observations used in Linear Regression Analysis to be smaller than expected. I attempted to solve this by applying less filtrations on the dataframe, which successfully increased the number of observations. Nonetheless, more data points with the same players and years would have helped me a lot during data merging. A way to improve next time is to collect more datasets with intersections and iterations to make my predictions closer to the actual target data. Another limitation was that I didn't consider a lot of unavailable data that could've actually mattered to salary calculations. These potential data columns were not included in the data research because they were not directly listed in the given csv files. Some of the possible factors that could have had an impact on players' salaries were cost of living, leadership, experience, and past teams. The last limitation was that I didn't consider non-linear or complex models, which could have modeled better regression models and made better predictions than linear regression. A way to improve is utilizing powerful models like machine learning to perform more accurate data analysis and prediction.

Citations

"Earned Run Average (ERA)." *Major League Baseball*, MLB Advanced Media, www.mlb.com/glossary/standard-stats/earned-run-average. Accessed 25 Nov. 2022.

"Runs Batted In (RBI)." *Major League Baseball*, MLB Advanced Media, www.mlb.com/glossary/standard-stats/runs-batted-in. Accessed 25 Nov. 2022.

"Baseball Databank." *Kaggle*, OPEN SOURCE SPORTS, 2019, www.kaggle.com/datasets/open-source-sports/baseball-databank. Accessed 25 Nov. 2022.